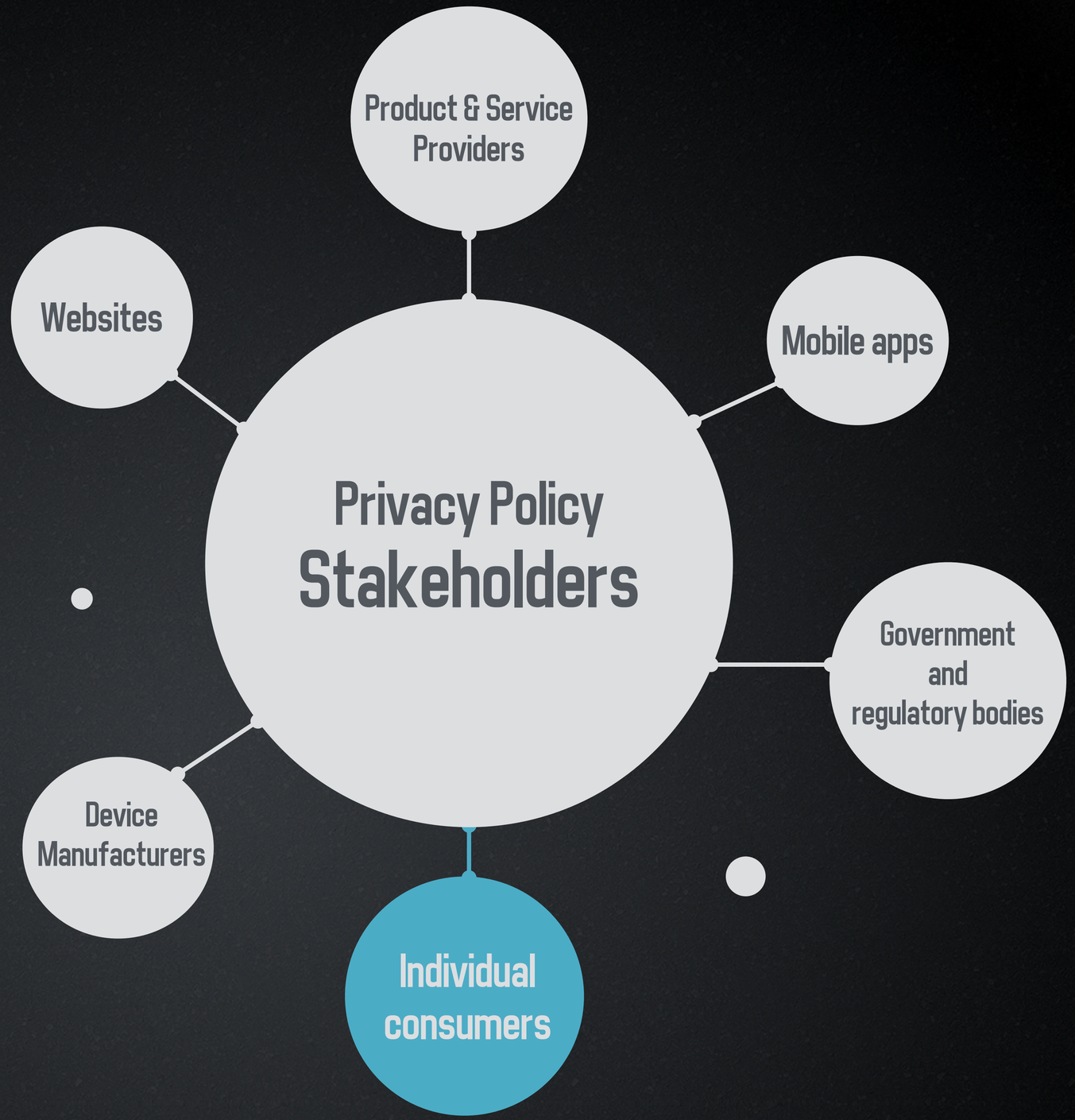
Detecting Textual Saliency in Privacy Policy

DS 5500
FALL 2021

Phase-1 Pitch

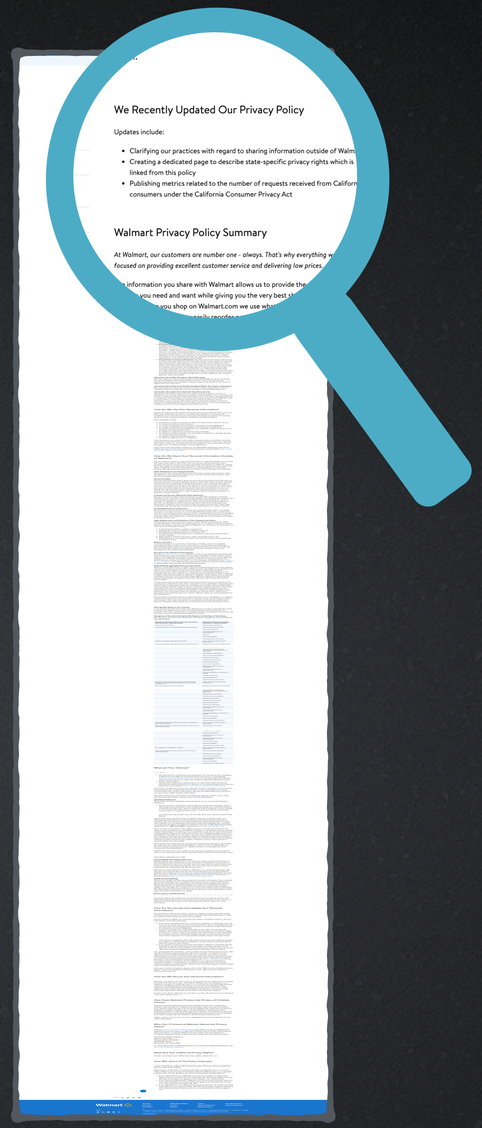
Kaushik Nishtala ■ Shubhanshu Gupta

Introduction



A Close Look At Privacy Policy

- Name
- Password
- Email
- Billing Data
- Profile Info
- SSN
  Account Info
- Browsing
  Activity
- Uploaded
  Media
- IP
  Address
- Physical
  Addresses
- Device IDs
  and Cookies
- Demographic
  Info
- Social Network
  Info
- Licences or
  Insurance
- Employment
  Data
- Other
  Info/Media



Details about

- ? How and why do they collect our information?
- ? Are they sharing this data with third parties?
- ? How long is my data stored?
- ? Can I access my data?
- ? How secure and protected is my data?

& more

Challenges

01 Increase in complexity

More than 85% of privacy policies scraped are at or above college level reading difficulty (Source)

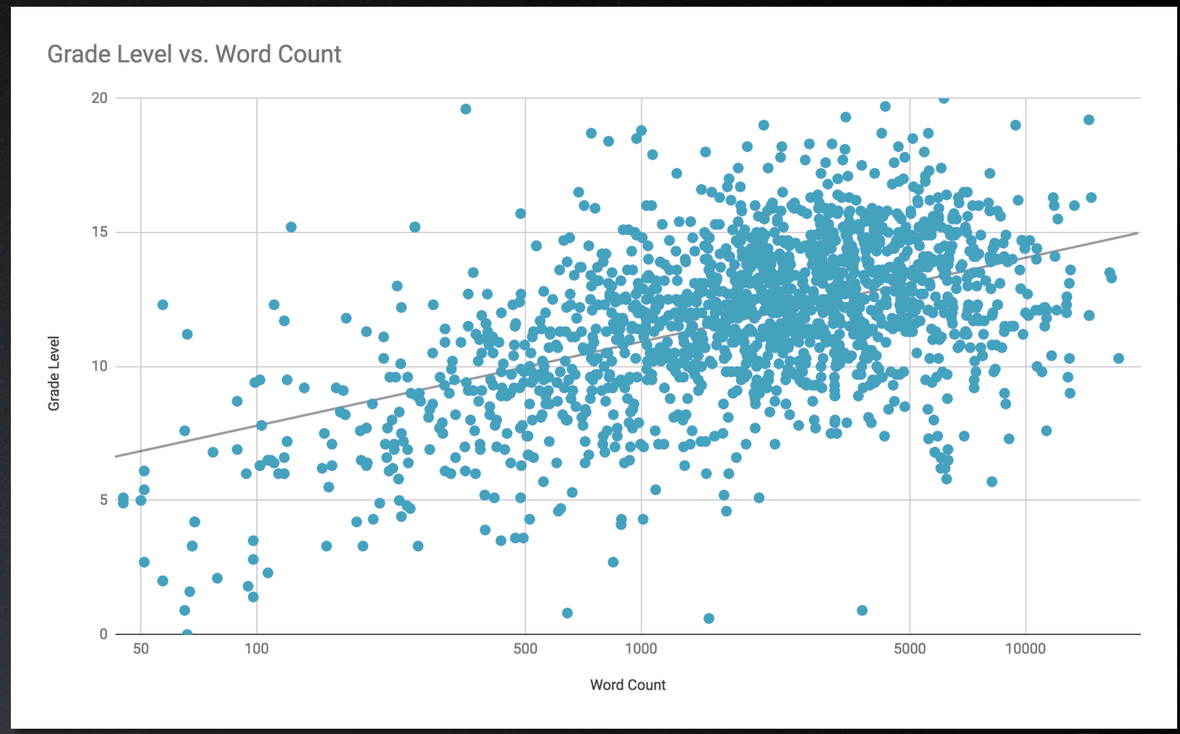
02 Increase in length and verbosity

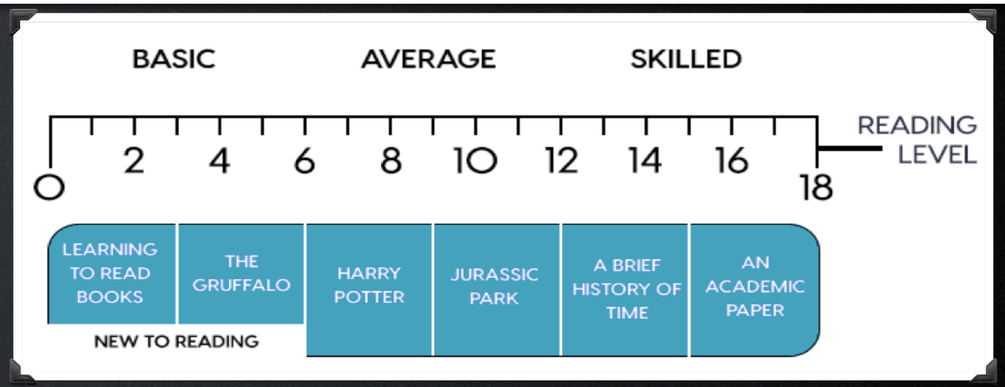
Privacy policies are 1.8x longer than in 2009 (Source)

Cause: Increase in data privacy regulations

Outcome: Opaqueness in privacy policies

Grade Level vs Word Count





Flesch-Kincaid Grade Level

(Source)

Goal

Provide key information to end-users using Machine Learning and Natural Language Processing

INFORM
Module

- ★ Visualization
- ★ Annotating policies

QUERY
Module

- ★ Enable users to ask policy-related

Interactive

Web Framework

INFORM

MODULE

Training Data

Online Privacy Policies (OPP-115), 2016

- ★ Corpus containing privacy policies of 115 US companies
- ★ Annotated by law students to specify data practices within the policy
- ★ An individual data practice in the corpus belongs to at least one of the ten data practice categories
- ★ Each category is further classified into a set of practice attributes.

Categories

01 First Party Collection/Use
02 Third Party Sharing/Collection
03 User Choice/Control
04 User Access/Edit & Deletion
05 Data Retention
06 Data Security
07 Policy Change
08 Do Not Track
09 International & Specific Audiences
10 Other

Other Datasets

MAPS Policies Dataset (PETS 2019)

- Consists of the URLs of 441,626 privacy policies.
- These were discovered as part of the Google Play Store app analysis conducted by the Mobile App Privacy System (MAPS) from April to May, 2018.
- Primarily used for creating domain specific word embeddings to fine tune CNN and BERT models

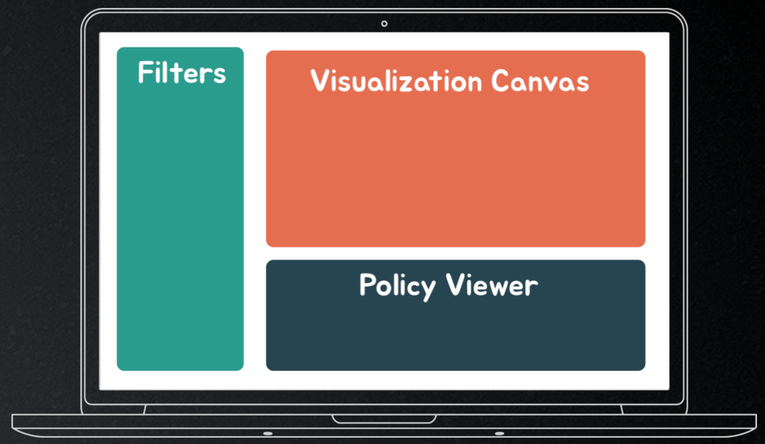
Princeton-Leuven Longitudinal Corpus of Privacy Policies

- Potential Test data for our trained models
- Consists of 910k privacy policies from130,000 websites across various industries
- Non-labeled longitudinal dataset spanning over two decades
- May use a policy scraper instead if the dataset is not procured by mid-phase

Visualization Framework

- ★ Convey high-level insights from the corpus to general consumers and/or regulators
- ★ Filter on metadata and attribute values to generate plots for each privacy category
- ★ The produced plots may answer holistic questions such as
  - ? How is my health/location information used?
  - ? What information collection can I opt-out from?
  - ? Why is my information shared?
  - ? What information is shared with external third parties?
  - ? What choices are available/not available?
  - ? To what uses of information do opt-outs apply?
  - ? Can I view/edit/delete/export my account?
  - ? How long is my information retained for advertising or marketing purposes?
  - ? How is my information protected?

OPP-115 Corpus + Metadata



Interactive
Web Framework

Powered by

Flask

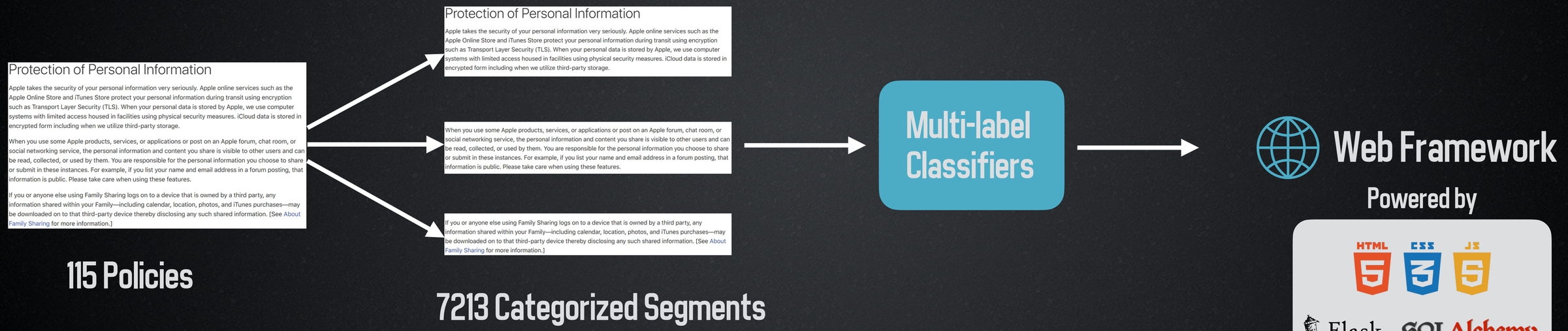
SQLAlchemy

HTML CSS JS

Dash
by plotly

Machine Annotation of Privacy Policies

- Predict one or more categories for each paragraph (segment) of a given privacy policy
- Train and tune models on OPP-115 corpus
- Test on either the Princeton longitudinal dataset/policies scraped using a web-crawler
- This segment-based classification is particularly appropriate because a segment can contain information about multiple categories, such as the first-party collection of data, third-party sharing, and data security.



Methods

Traditional Models

- ★ Build multiple category-specific classifiers that accept word vectors transformed using Paragraph2Vec and the GENSIM toolkit
- ★ Feed the word vectors to Classifiers such as Logistic Regression, Support Vector Machines etc
- ★ Explore a sequence labeling approach to apply hidden Markov models (HMMs) to privacy policy text.

Language Neural Models

- ★ Exploit MAPS corpus of 441k policies to train custom word embeddings for the privacy policy domain using FastText
- ★ Build Convolutional Neural Networks with general-purpose pre-trained embeddings such as Word2Vec, and GloVe

  CNN w/ Word2Vec | CNN w/ GloVe | CNN with domain specific embeddings

- ★ Evaluate Bidirectional Encoder Representations from Transformers (BERT) with the following variants:

  BERT-base | Fine-tuned BERT with domain specific embeddings | LEGAL-BERT

Powered by

Hugging Face

Google Cloud Platform

PyTorch

QUERY
MODULE

Freeform Question-Answering System

- ★ Frame the QA task - Information Retrieval QA or NLP QA
- ★ Exploit a recently released PolicyQA dataset curated from the OPP-115 corpus
  - ◆ Enables us to model the task as predicting the answer in the given policy segment using existing neural approaches from literature (utilizing Hugging Face transformer API)
  - ◆ A similar dataset SQuAD (Stanford Question Answering Dataset) curated from Wikipedia articles could help provide more insight into the process.
  - ◆ Utilize state-of-the-art architectures such as Bi-directional Attention Flow (BiDAF) network or BERT with their variants as before

Powered by

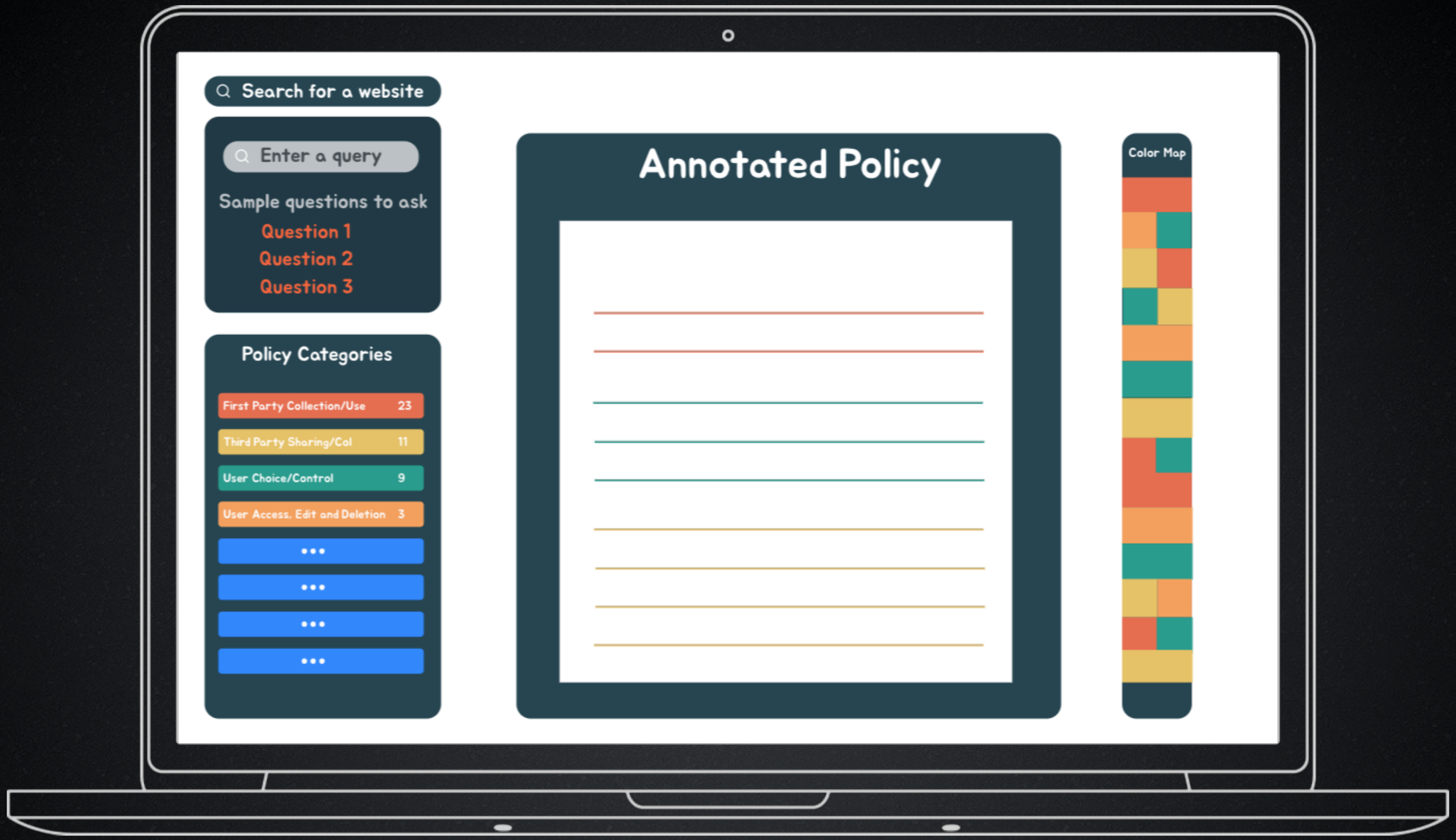
Hugging Face

Google Cloud Platform

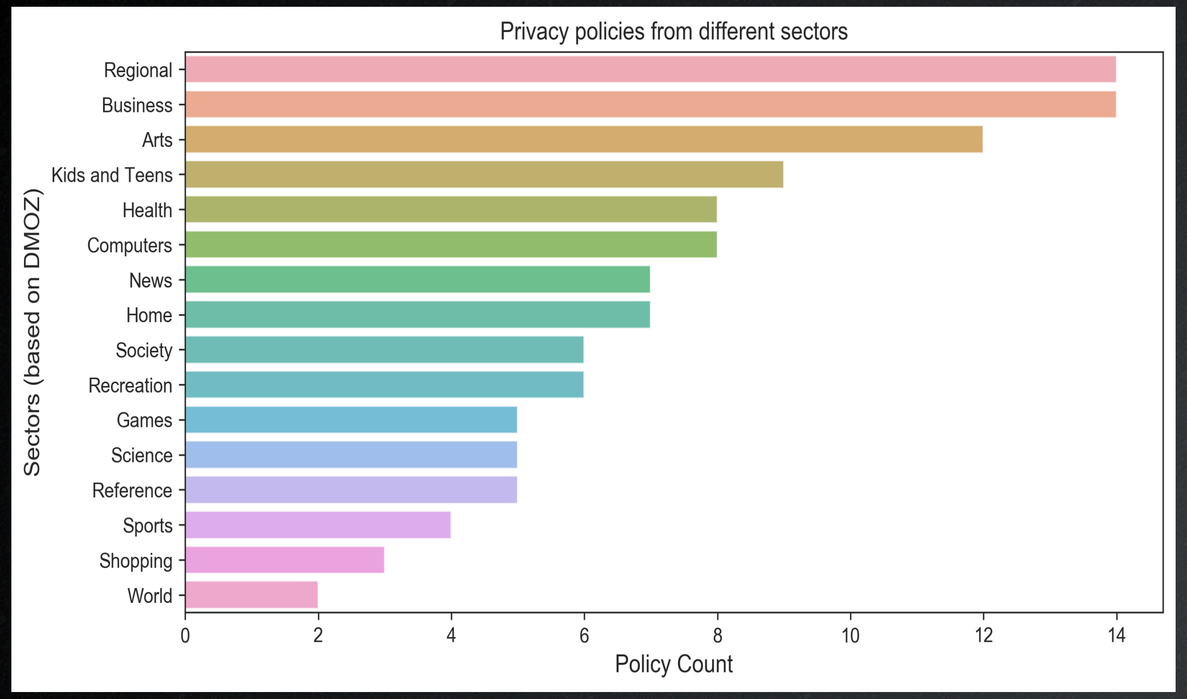
PyTorch

	Example 1	Example 2
Policy Text	Information You Give Us: We receive and store any information you enter on our Web site or give us in any other way. Click here to see ...	Promotional Offers: Sometimes we send offers to selected groups of Amazon.com customers on behalf of other businesses. When we do this, we do not give that business your name and address. If you do not want to receive such offers, ...
Question	How do you collect my information?	Is my information shared with others?
Answer	information you enter on our Web site	we do not give that business your name and address

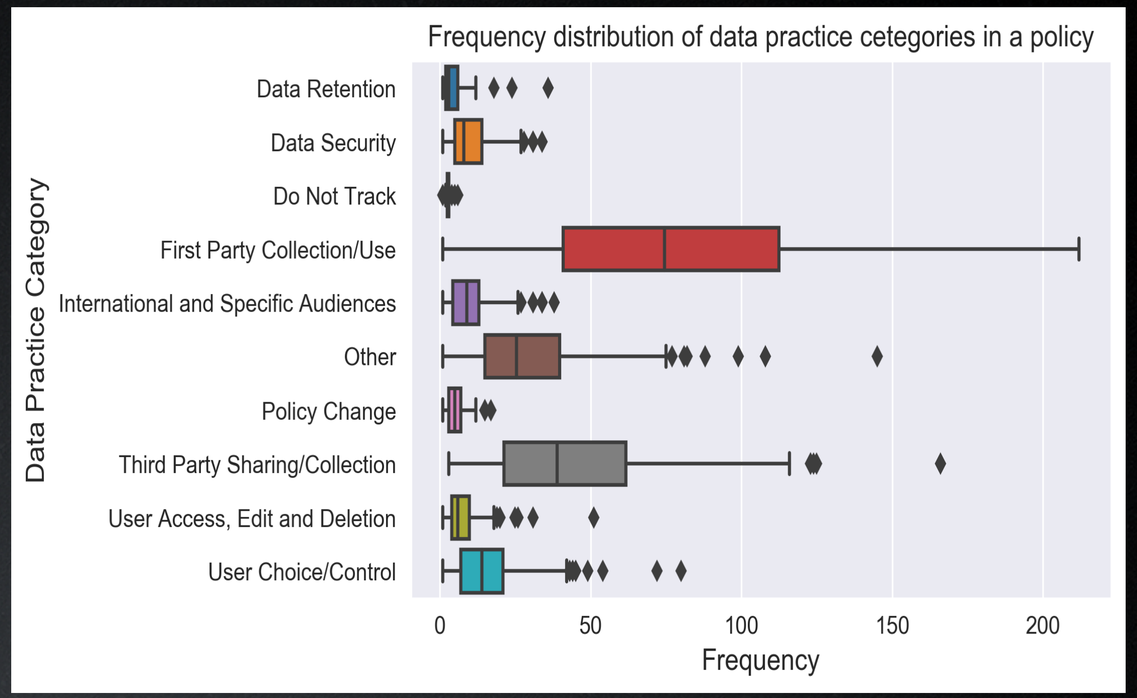
Machine Annotation & Query (Website Mockup)



Preliminary Results

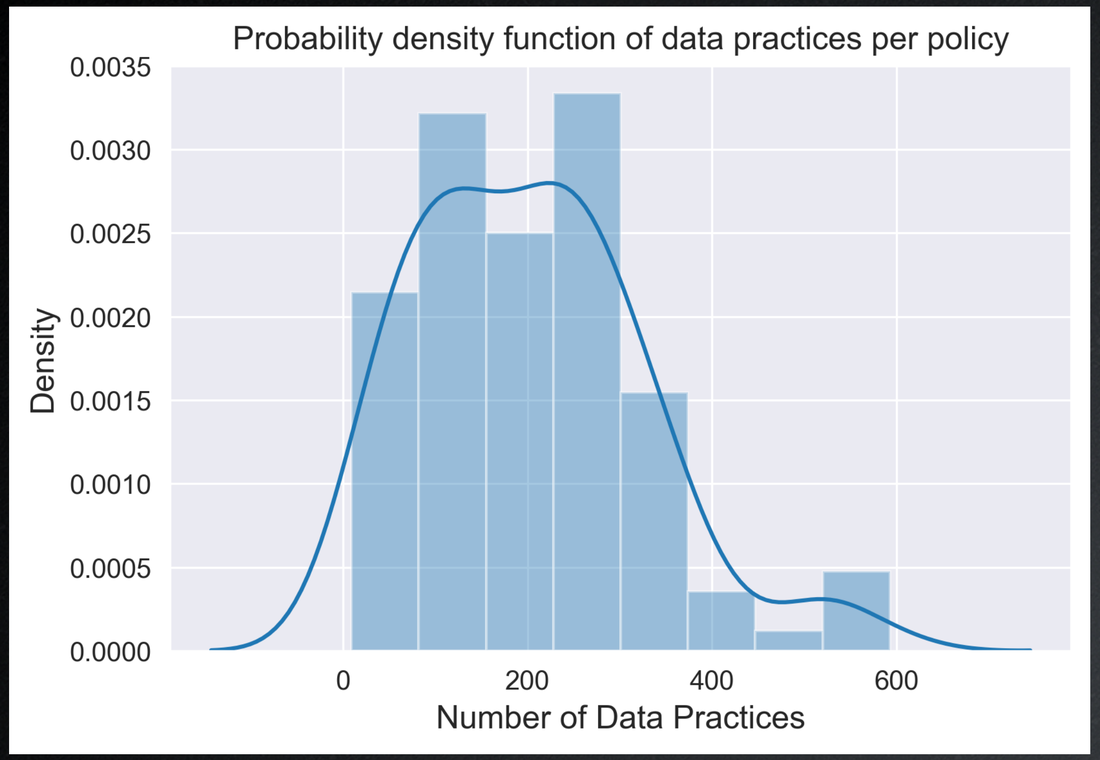


Distribution of the sectors (based on DMOZ) in the corpus

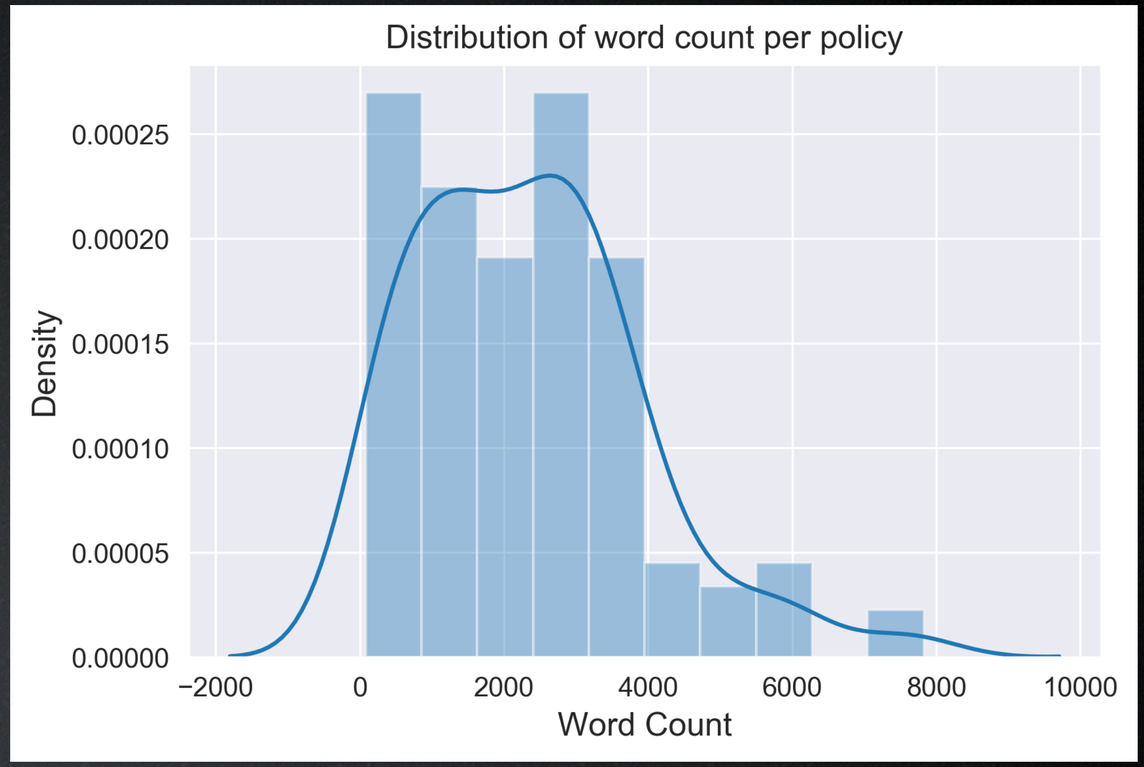


Frequency distribution of data practice categories in policies

Preliminary Results (Continued)



PDF of number of data practices per policy



Word count distribution per policy

Project Timeline

Full-semester project

- ☆ Literature Survey
- ☆ Preliminary Results

01 Phase 1 Pitch

- ☆ Data Preparation
- ☆ Visualizations
- ☆ Traditional models
- ☆ Front-end (some progress)

02 Phase 1 Progress Updates

- ☆ Implement Front-end + Back-end (viz)
- ☆ Language Neural Models (except BERT)

03 Phase 1 Presentation

- ☆ Transformer models
- ☆ Literature Survey (QA systems)

04 Phase 2 Pitch

- ☆ QA using Transformers
- ☆ Front-end + Back-end

05 Phase 2 Progress Updates

06 Phase 2 Presentation

INFORM module — QUERY module

Potential Challenges

- If and how well do our models work without a specifically trained BERT?
- Stay focused and diligent with the timeline - as the project scope is wide
- As the labeled data is from 2016 and earlier, how well does it generalize to 2020+?
- Choosing Multi-label Classification Evaluation Metrics (Micro vs Macro averages) with imbalanced data

THANK YOU